ENGINEERING-PSYCHOLOGY RESEARCH LABORATORY

University of Illinois at Urbana-Champaign

TECHNICAL REPORT EPL-84-1/NASA-84-1

FEBRUARY 1984

An Investigation of the Dissociation Between Subjective Measures of Mental Workload and Performance

Yei-Yu Yeh

Christopher D. Wickens

Prepared for

National Aeronautics and Space Administration
NASA Ames Research Center
Moffett Field, CA
Contract No. NASA-NA-G-2-169

An Investigation of the Dissociation Between Subjective Measures of Mental Workload and Performance

Yei-Yu Yeh and Christopher D. Wickens

Abstract

This report describes research conducted during the first years under a contract from NASA Ames Research Center; Dr. Sandra Hart was the technical monitor. The report addresses the dissociation between subjective measures of mental workload and performance. Three generic factors are identified that will drive subjective workload upward more than driving performance downward: Perceptual (versus response) load, an increased number of tasks, and better data quality. One factor, resource competition, is assumed to drive performance more than subjective workload. The theory of dissociation is tested in three experiments that employ different variations and combinations of three different tasks (tracking, memory search, and a simulated air traffic control task). The predictions of the theory are generally supported by the data. In addition, various subjective scales of mental workload are tested across the experiments. The correlations between these scales and multi-dimensional scaling data are used to help interpret the hidden cognitive structure of task difficulty.

Introduction

The system designer makes a modification of an existing system, intended to improve performance. Operators who use the system with its new modification are unanimous in preferring it to the old prototype. Yet performance on the system is clearly worse than with the original version. Which version is "better?" This is an example of a dissociation between workload measures. The following paper is intended to present and test a theory of dissociation to interpret their source and examine their implications.

Performance and subjective ratings have both been considered to be sensitive workload measures (Casali & Wierwille, 1982; Hicks & Wierwille, 1982; Rahimi & Wierwille, 1982; Wierwille & Gutmann, 1978; Wierwille & Connor, 1983). The view that both measures are functionally equivalent means of measuring mental workload seems to be supported by a high correlation between subjective and performance measures across a wide range of tasks and task configurations. Tasks that are performed more poorly are generally described to be more difficult (see Moray, 1982 for a review). However, dissociation has been cited in both laboratory (Derrick, 1981; Wickens & Derrick, 1981; Wickens & Yeh, 1982, 1983) and in a more real world environment (Herron, 1981).

The dissociation phenomena may simply result from the difference in the natures of the two measures. While little is known about the origin of the feeling of subjective workload (Moray, 1982), the nature of human performance has been thoroughly examined. Wickens (1980, 1984) concluded that the human information processing system is composed of separate resources of limited quantity. Three dichotomous dimensions are important for defining the multidimensional structure of processing resources. These three dimensions are defined by: (1) the auditory vs.

visual modality, (2) visual vs. spatial processing codes, and (3) perceptual cognitive vs. response-related information processing stages.

To examine how subjective workload relates to the underlying nature of human performance, Derrick (1981, 1984; Wickens & Derrick, 1981) investigated how the two measures change as the task difficulty is manipulated by resource cost. Resource cost was manipulated via different patterns in the three dichotomous dimensions as defined in the multiple resource model, i.e., input/output modality, stages, and processing code.

The results showed that performance in contrast to subjective workload is driven relatively more by the difficulty of a single task while subjective workload is driven relatively more by the number of the task performed at the same time. In this study, the similarity of task difficulty was also rated for all possible pairs of task combinations. These similarity data were analyzed by a Multidimensional Scaling (MDS) method i.e., INDSCAL to disclose the psychological structure of task difficulty. The result from this analysis indicated that the underlying structure can be partially explained by task resource competition where resources are defined by stages, codes, and input/output modality.

The correspondence between the psychological structure of task difficulty and the dimensions of resources supported the view that subjective workload can be understood by empirically supported constructs in the multiple resource model. However, the two measures may be differently affected by the differing demands on the resource pools. Human performance is determined by the interaction of the capacities of a large number of different subsystems and the demand imposed upon those subsystems. However, demand on some subsystems may not be precisely "read" when the operator generates an introspective rating of mental workload. Dissociation would be revealed when certain subsystems contribute heavily to one measure but not to another. For example, the number of resources required in processing appears to be weighted heavily in subjective workload estimates but only marginally contributes to performance. Thus, when dual tasks impose demand on several resource pools, they are performed considerablybetter than their subjective ratings would indicate. On the other hand, the difficulty of a single task may contribute more to performance than it does to subjective workload. In this case subjective reports would provide an overly optimistic view of the expected system demands.

Why are processing characteristics "read" differently in the two measures? A model of introspective verbal reports offered by Ericsson and Simon (1980) provides some clues. According to this model, subjective introspection reflects information heeded in working memory. Any information to be verbally reported has to be in working memory. Processes that do not utilize capacity in working memory such as automatic processing will not be available to introspection. Processes whose demand exceed the maximum capacity of the memory will not be accurately reported because there is less variation in resource mobilization under such a condition.

Subjective workload estimates are the introspection of mental workload involved in performing the tasks. Thus, they are verbal report data and may primarily reflect information in working memory. This postulation conform with evidence from Derrick's study. Increasing the number of tasks demand more resources in working memory to coordinate and execute the processing of each task. Therefore, subjects feel more loaded when they must perform two tasks at the same time than when only one task is required. In studies adopting the task-characteristic approach, the variables that change subjective workload are also related to the demands on the working memory. These variables include memory load (Hauser, Childress, & Hart, 1982; Eggmeier, Crabtree, & Reid, 1982), rate of presentation (Hauser et. al., 1980), generating lead in a second order tracking task (Ashkenas, 1966), number of decision alternatives (Borg, 1978), insufficient data (Borg, 1978), fraction of attention (Hess, 1977), number of tasks to-be-processed per processing unit (Tulga & Sheridan, 1980) etc. When the demand in working memory increases by higher memory load, fast presentation rate, generating lead, more decision alternatives to choose, or more tasks to be processed per processing unit, subjective workload increases. Futhermore, Eggmeier et. al. (1982) found that subjective ratings are not sensitive as performance measures when memory load is too high. This finding also confirm Ericsson and Simon's assertion that verbal report cannot tap the minute variation in a high memory load condition.

A similar view has been presented by Gopher and Braune (1982). Subjecive measures are assumed to reflect the perceived magnitude of resource investment in the conscious attention. They concluded that subjective estimates follow the pattern of the most resticted model of a single undifferentiated pool of resources. In other words, subjective estimates only reflect part of the information processing while performance follows the pattern of a multiple resource model.

In the multiple resource model, working memory is represented as dependent upon the perceptual/central resources. Hence, subjective workload is suggested to be dominated by the demands upon these resources, and will be less sensitive to demands imposed upon the resources related to responding. Based upon this assumption, a tentative theory of the factors that drive subjective workload (S) and performance (P) is presented in Table 1.

Table 1. Sources of Performance - Subjective Dissociation

Source	P	S
1. Increased Single Task Difficulty	2	1
Perceptual/Cognitive	2	1
Response	2	0
2. Concurrent Task Demand	1	2
Same Resources	1	2
Different Resources	0	2
3. Motivation		
(Resource Investment)	-1	+1

A set of manipulation or sources are listed in the first column. These represent things that can be done to a task to increase general workload (decrease performance and/or increase the subjective feeling of effort). Within the second column is a number that indicated the extent to which the manipulation will deteriorate performance (P). The number within the third column indicates the extent to which the manipulation will increase the subjective difficulty (S) of the task. The important factor governing dissociations is the ratio or relative value of these two numbers for a given source. The particular numbers have ordinal interpretations only.

The theory predicts that manipulating the parameters of a single task will generally influence P more than S (Wickens & Derrick, 1981). This difference is particularly pronounced if the task is degraded by imposing demands on responding (affecting S with a 0 value); rather than on perceptual/cognitive processing, affecting S with unity value.

Below the solid line, the theory predicts that increasing workload by increasing the number of tasks that must be performed concurrently will generally serve to increase S and decrease P, but the former will be varied by a greater degree than the latter (Wickens & Derrick, 1981; Wickens & Yeh, 1982). Futhermore, the subjective experience of workload is uninfluenced by whether those tasks compete for common or separate processing resources within the multiple-resource system. However, the defining property of the multiple resource theory of dual task performance asserts that performance will be unaffected to the extent that tasks compete for separate resources, but will degrade when common resources are employed. A final source of dissociation is reflected in item 3 of the table. Here we predict that any variable which induces the investment of more resources into a task either through greater motivation, or by providing better quality data upon which to operate will simultaneously improve performance while leading to the feeling of greater mental workload. An example might be that of increasing the gain on a tracking display, or the observation that increased incentive will produce these effects (Vidulich & Wickens, 1983). On the other hand, when subjects lower their motivation and performance criterion, their performance will deteriorate while they feel less loaded (Tulga, 1978).

Methodology and General Experimental Paradigm

Methodology. In the three experiments we report below, we examined the relative effects on mental workload and performance of a series of manipulations on three prototypical tasks. Any manipulation can be characterized by a vector in a two-dimensional space defined by the changes of S and P that occur as result of the manipulation. The orientation of this vector (the ratio of $\Delta P/\Delta S$) describes the relative impact of the manipulation on the two variables. In theory, then it should be possible to define a dissociation in a manipulation to occur whenever $\Delta P/\Delta S = 1$. However, our current approach is to examine the dissociation between pairs of manipulations, such that the vector orientation ($\Delta P/\Delta S$) is different between the two. In order to test for this dissociation statistically, we first convert each change in P and S into standard scores by dividing each by measures of its variability across the conditions investigated. Hence, each measure now spans a range of comparable units. Measure type (S vs. P) is then treated as one factor in ANOVA, with the nature of manipulations representing the second multi-level factor. A dissociation is verifid to exist when there is an interaction between type and the kind of manipulation, indicating that the vectors point in different directions.

General Experimental Paradigm

Three experiments were conducted. Three tasks were employed across the three experiments, with various manipulations of difficulty possible in each. The three tasks are as follows:

(1) Compensatory tracking. In this task subjects tracked a random disturbance input with a bandwidth of either 0.32 Hz (easy) or 0.54 Hz (difficult). The task could be performed either singly (the right hand tracking vertically, or the left horizontally), or in a dual axis combination of the two single task configurations. Finally, the second order task could be performed either with or without a predictor symbol, driven by the estimate of the cursor's current velocity and acceleration.

(2) Memory search task. Prior to each trial, subjects viewed a set of three 3-item, alphanumeric strings of the form A29, J93, and M46. These strings were then held in working memory for the next two minutes as the subject was presented a series of probes, half of which were in the memory set. Subjects were to respond as rapidly as possible if the probe

was or was not contained in the memory set. In various configurations the stimuli were presented either auditorily (A) or visually (V) and subjects could respond either manually (M) or with speech (S). Stimuli occurred at irregular intervals ranging between 3 and 5 seconds. The task could be performed by itself, or in combination with either of the other two tasks.

(3) "Crash" task. In this simulated air traffic control task, the subject viewed two aircrafts, closing on an approach to each other, as shown in Figure 1. The subject's task was to make two decisions, as rapidly as possible in the following sequential order. (1) Which plane would pass in front of the other, and (2) At the point at which the trailing plane will be directly abeam of the leading plane, how close will the two planes be? The second decision required a category rating from 1 to 5. In different conditions the response was either entered manually or spoken. Performance was scored on the basis of the latency of both decisions, the accuracy of the first decision and the correlation between actual and judged proximity for the second. Two levels of difficulty were created by creating a set of problems with actual separations of 1, 3, and 5 display units (easy) and 1, 2, and 3 (difficult). Problems were presented on a self-paced schedule with each new problem occurring from 1 to 3 seconds after the preceeding one. This resulted in an average of one problem about every 5 seconds.

The three tasks were chosen to meet two criteria: (1) Each is representative of the kinds of activities performed by pilots, (2) as in previous studies (Wickens, Mountford, & Schreiner, 1981; Derrick, 1981), each is designed systematically to place demands upon various components of processing resources within the framework of the multiple resource theory (Wickens, 1984). The three tasks with their various input/output configurations as well as the particular resource loaded by the task

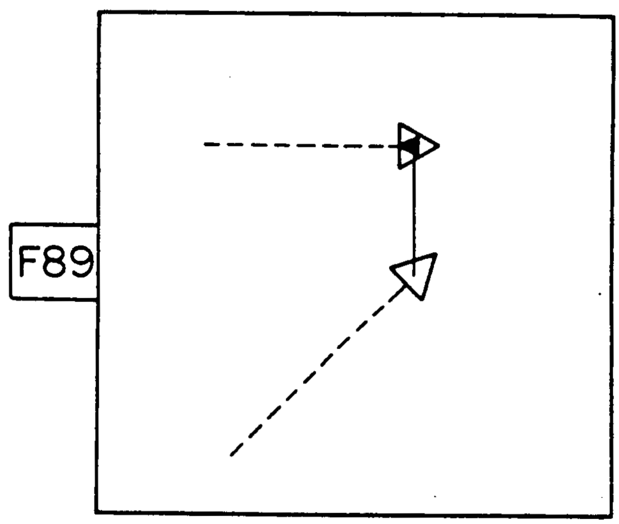


Figure 1: Stimuli for the Crash Task.

within each dimensions of the multiple resource model are shown in Table 2.

Table 2. Processing Resource Demands of the Tasks

	Task		
Resource Dimension	Tracking	Memory	Crash
1. Stage Defined	Response	Perceptual/ Cognitive	Perceptual/ Cognitive
2. Code Defined	Spatial	Verbal	Spatial
3. Modality Defined	Visual-Manual	Visual-Manual Auditory-Speech	Visual-Manual Auditory-Speech

In each experiment, subjects practiced the tasks extensively over a period of three days. These were followed by three days of experimental data collection, during which each task configuration was rated both in terms of its overall level of mental workload and on a series of other subjective scales. Subjects were also asked to rate the similarity of task difficulty in all possible pairs of task combination. These data were subjected to multidimensional scaling analysis to identify the cognitive structure of task difficulty. Ratings of all unidimensional attributes were correlated with the disclosed structure to aid the interpretation of each axis. At the end of the last experiment, subjects were also asked to rate the importance of 19 factors to the way they define their mental workload. Results of these ratings were compared to the results obtained by Hart, Childress, & Hauser (1982).

Experiment 1

In this experiment, the difficulty of a compensatory tracking task was manipulated by three means: (1) changing the difficulty of the single task, (2) adding a concurrent tracking task which competes for the same resources, and (3) adding a Sternberg memory search task employing auditory input and speech response which thereby demands separate resources from the tracking task. The purpose of this study is to replicate Derrick's (1981) finding and to verify the prediction that P is driven more by the difficulty of a single task, and S is driven more by the number of concurrent tasks.

Method

Subjects. Eight right-handed male students of University of Illinois were paid volunteers. All subjects were native speaker of English and have correct vision.

Sternberg memory search task. Stimuli for the task were always presented auditorily and responded by speech (AS).

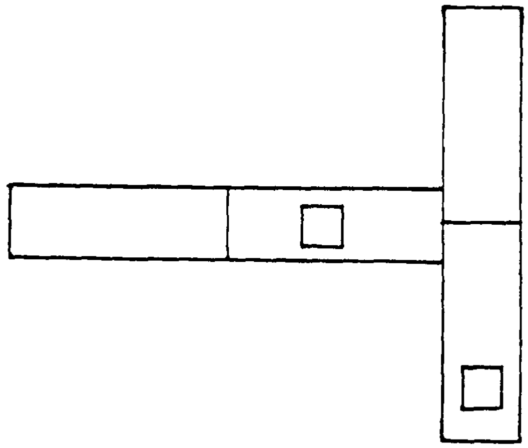
Tracking task. The difficulty of the tracking was varied from the easy condition either by increasing the control order, or the bandwidth. The control order was increased by varying the proportion of first and second order component from 0.0 to 0.5 to 1.0 The bandwidth was increased by changing the upper cutoff frequency to a value of 0.54 Hz. The second order tracking task could be performed either with or without a predictor. These five single tracking tasks (i.e., first order with low bandwidth, first order with high bandwidth, second-order with a predictor, second-order without a predictor, and mixed order) were performed either with the left or the right hand. Thus, 10 single tracking tasks were performed by the subjects.

Dual task. Three dual-task conditions were performed. An easy tracking task was paired with each of the following three tasks: (1) the Sternberg memory search task in AS input/output modality, (2) another tracking task in a separated display (Figure 2a), and (3) another tracking task in an integrated display. In this integrated display, the X and Y errors were displayed by the position of a horizontal and a vertical line, respectively (Figure 2b). The subject's task was to keep the intersection of the two lines on a reference cross in the center of the display. Subjects always tracked horizontally with the left hand and vertically with the right hand. In all three dual task conditions, subjects were instructed to treat the two tasks with equal priority.

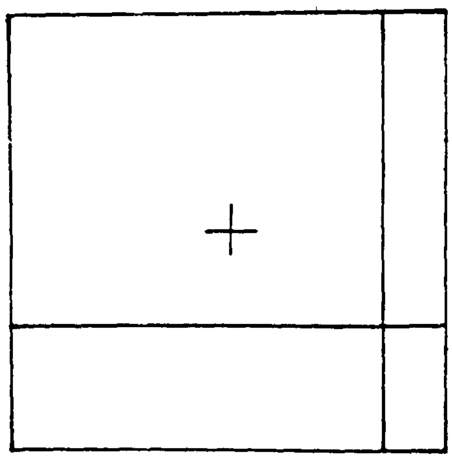
Procedure. All subjects performed in six 1-2 hour sessions. Subjects practiced all the tasks extensively over a period of three days. (All of the conditions are presented in Table 3). These were followed by three days of experimental data collection (sessions 4,5, and 6). In the fourth session, subjects performed the baseline condition (i.e., first-order with a low bandwidth) first and were told to treat it as the standard task. They then performed all other tasks in a random order and rated each one against the standard task on four attributes: overall workload, task complexity, psychological stress, and time-demand.

A value was assigned to the standard task (5, 10, or 20 was randomly chosen) and subjects were instructed to give a value to the just-performed task on each attribute. For example, subjects were asked "Assuming that the overall workload in doing the standard task is 20, what is the overall workload in performing the task just performed? If it is twice the workload of the standard task, then assign 40 to this task. If it is about half of the workload of the standard task, then you just give a value of 10. Assign the workload value with the standard task as the reference."

At the end of the session, subjects were presented a list of all tasks in a random order. They were asked to do the ratio ratings on eight



(a) Dual-axis tracking in a separated display



(b) Dual-axis tracking in an integrated display

Figure 2: Dual-axis tracking displays in Experiment 1

I. Single task conditions

(1) low bandwidth 1st-order (α=0.0) tracking on the left hand.
(2) high bandwidth 1st-order tracking on the left hand.
(3) mixed 1st- and 2nd-order tracking on the left hand.
(4) 2nd-order tracking without a predictor on the left hand.
(5) 2nd-order tracking with a predictor on the left hand.
(6) - (10) same conditions as above tracked on the right hand.
(11) AS memory search task.

II. Dual task conditions

(12) Dual-axis tracking tasks with a separated display
(13) Dual-axis tracking tasks with an integrated display
(14) Tracking with an AS memory search task

Table 3 - Experimental conditions (Experiment 1)

attributes (i.e., the four attributes rated at the end of a trial along with four additional attributes: mental effort, task complexity, response complexity, and task demand). They were also asked to rate each task on a 9-point scale of each attribute, and to rank order all tasks by their difficulty. In addition, 20 pairs of task combinations were randomly selected. These pairs were presented to subjects and they were instructed to rate the similarity of task difficulty in each of the pairs (Derrick, 1982). They were told that they could use any attribute(s) to make the judgment but they should use the same ones to judge every pair.

During the fifth session, subjects performed and rated the tasks in the same ways as they did in session 4. In addition to the eight attributes, four more attributes were rated at the end of this session. These four were feedback adequacy, success of performance, task nature (a modified Cooper-Harper scale), and the excess capacity by doing the task (see Table 4 for all scales used in the experiment). Similarity judgments were not rated in this session.

In the last session, subjects performed all of the tasks. The only subjective ratings required in this session were the similarity judgments of difficulty on all possible pairs of task combination.

Results

Correlational Analysis

(A) Test-retest reliability. The test-retest reliability of subjective ratings on each task condition was computed between sessions 4 and 5. These data are shown in Table 5. The ratio ratings in the fourth session were Log transformed and correlated with the corresponding transformed ratings in the fifth session. The Pearson-product correlation was calculated for each attribute and for each subject to see how consistent (reliable) each subject was on rating tasks on each attribute.

Scales	(Description)
Overall Workload	Overall workload of doing the task.
Complexity	How complex was the task?
Psychological stress	The psychological stress experienced in doing the task.
Time-demand	How busy were you in doing the task?
Task demand	How demanding was the task?
Input Complexity	The complexity of input stimuli.
Mental Effort	How much mental effort did you make in doing the task?
Response Complexity	The complexity of responding the task.
Rank order	Rank order the task difficulty.
Feedback adequacy	How adequate was the feedback you received in doing the task?
Success of perfermance	How successful were you in doing the task?
Nature of the task	How do you think of the task? Choose one of descriptions: (1) very easy to do with excellent precision (2) very easy to do with good precision (3) easy to do with fair precision (4) doable with somewhat inadequate precision (5) doable, but only very imprecisely (6) difficult to do (7) very difficult to do (8) nearly doable (9) undoable
Excess Capacity	Assume all the capacity that you have is 100. How much is left by performing the task?

Table 4 - Rating scales

Scale	r Lg(ratio ratings) at the end of a trial	r Lg(ratio ratings) at the end of a session	r Interval ratings	r between the two methods
Overall workload	0.71	0.83	0.77	0.90
Complexity	0.83	0.89	0.89	0.92
Psychological stress	0.61	0.85	0.84	0.91
Time-demand	0.79	0.83	0.88	0.90
Task demand		0.85	0.88	0.88
Input complexity		0.82	0.86	0.91
Mental effort		0.85	0.80	0.86
Response complexity		0.85	0.84	0.89
Rank order			0.89	

(No of task = 14, all correlations are significantly greater than 0.0, $p < 0.05$)

Table 5 - Test-retest reliability on rating (Experiment 1)

These correlations were computed for ratings collected at the end of a trial and for ratings done at the end of a session. Generally speaking, most subjects were more consistent on rating tasks at the end of a session than at the end of a trial (Table 5). The overall reliability across subjects over four scales collected at the end of a trial was 0.73. The overall reliability of corresponding four ratings done at the end of a session was 0.85. The overall reliability over the eight ratio and interval scales at the end of a session were 0.85. The correlation between the transformed ratio and interval ratings was 0.90. The averaged reliability on the smilarity judgments of the 20 pairs collected at both session 4 and 5 was 0.79.

(B) Correlations among unidimensional ratings and with performance. End-of-Session ratings on each scale were averaged over all subjects, correlated with ratings on other scales, and correlated with the averaged tracking performance (Table 6). The inter-scale correlations were very high. These high correlations suggest that mental workload is related to every aspect of performing a task such as the complexity, the psychological stress, the input complexity, and so on. The correlations between mean tracking performance and mean ratings on each scale follow predictions. The correlation between the tracking performance and workload ratings was 0.74. This correlation suggests, not surprisingly that in a general sense tasks that are performed worse are viewed as more difficult.

Performance and Subjective Workload Analysis

The raw performance scores and the overall subjective workload in each task configuration are shown in Table 7. Tracking performance is found to decline monotonically across the 5 increasing levels of order. All other manipulations of the tracking task also increase its error and

	Wld	Comp	Strs	Time	TkDe	Inp	MnEf	Resp	Fb	Perf	Nat	Rank	Exce
Comp	.98												
Strs	.99	.97											
Time	.99	.98	.98										
TkDe	.99	.98	.99	.98									
Inp	.97	.96	.98	.99	.97								
MnEf	.96	.91	.95	.97	.95	.97							
Resp	.98	.99	.97	.97	.98	.96	.90						
Fb	.74	.69	.80	.71	.77	.74	.72	.71					
Perf	.85	.85	.89	.81	.88	.81	.76	.87	.73				
Nat	.90	.93	.92	.87	.92	.87	.78	.94	.71	.94			
Rank	.97	.96	.97	.95	.98	.95	.91	.97	.88	.91	.93		
Exce	.98	.98	.97	.97	.99	.95	.93	.98	.69	.88	.91	.97	
MDS dimensions													
D1	.96	.98	.94	.94	.97	.92	.90	.95	.81	.92	.91	.97	.95
D2	.63-	.63-	.72	.69-	.66-	.70-	.59	.70	.19-	.43-	.62	.63	.67-
D3	.12-	.12-	.02-	.20-	.08-	.30-	.33-	.02-	.16-	.17-	.15-	.03-	.02-
Tracking Error													
Rms	.74	.73	.80	.69	.78	.73	.68	.76	.74	.94	.89	.80	.75

Wld - Overall Workload
Comp - Complexity
Strs - Psychological stress
Time - Time demand
TkDe - Task demand
Inp - Input complexity
MnEf - Mental Effort
Resp - Response comlexity
Fb - Feedback adequacy
Perf - Success of performance
Nat - Nature of the task
Rank - Rank order task difficulty
Exce - Excess capacity
D1 - 1st psychological dimension
D2 - 2nd psychological dimenson
D3 - 3rd psychological dimension
Rms - Tracking error

(All correlations are significant at 0.05 except the ones with "-")

Table 6 - Correlation among unidimensional ratings , Performance, and MDS dimensions (Experiment 1)

	RMS	Accuracy	RT	Workload
1st-order	0.156			2.641
mixed order	0.184			3.719
2nd-order without a predictor	0.236			5.281
2nd-order with a predictor	0.183			4.969
High bandwidth	0.297			5.938
Dual-axis tracking (separated)	0.205			6.156
Dual-axis tracking (integrated)	0.215			6.438
AS memory task only		97.368	1.098	3.125
Tracking + AS memory task	0.177	97.105	1.050	5.219

Table 7 - Original scores (Experiment 1)

its perceived difficulty as well, although to various degrees. This difference in the degree of increase in S and P impacted by the different manipulations will be subject of the following analysis of dissociation.

Performace of the standard task (i.e., low-bandwidth first order tracking) was used to compute the performance decrements for all other tracking tasks. Performance of the single Sternberg task was used to compute the decrement imposed by performing the same task in the dual-task condition. For the dual task tracking and Sternberg combinations it was necessary to derive a single integrated measure of the performance decrement of both tasks. This was accomplished in the following fashion: The decrement scores were calculated for each dependent variable in the way suggested by Wickens, Mountford, and Schreiner (1980). The variability of the dependent variable in each condition over the last three sessions was obtained. The variabilities in all single tasks were averaged as well as the variabilities in all dual-task conditions. An overall measure of variability from the two (i.e., single and dual-task conditions) was averaged for each subject as a measure of the mean within-subjects variability. This quantity was then used as the normalizing factor for each dependent variable. Normalized decrements on the left hand task performance were then averaged with normalized decrements on the right hand performance to indicate the performance changes for each dual task manipulation.

Changes in ratings of overall workload from the standard were calculated and normalized in the same way as performance. Consequently, both performance and subjective ratings of workload were scaled on equivalent units before further analysis.

The normalized performance and workload ratings produced by each manipulation are presented in Figure 3a (these data do not include those of

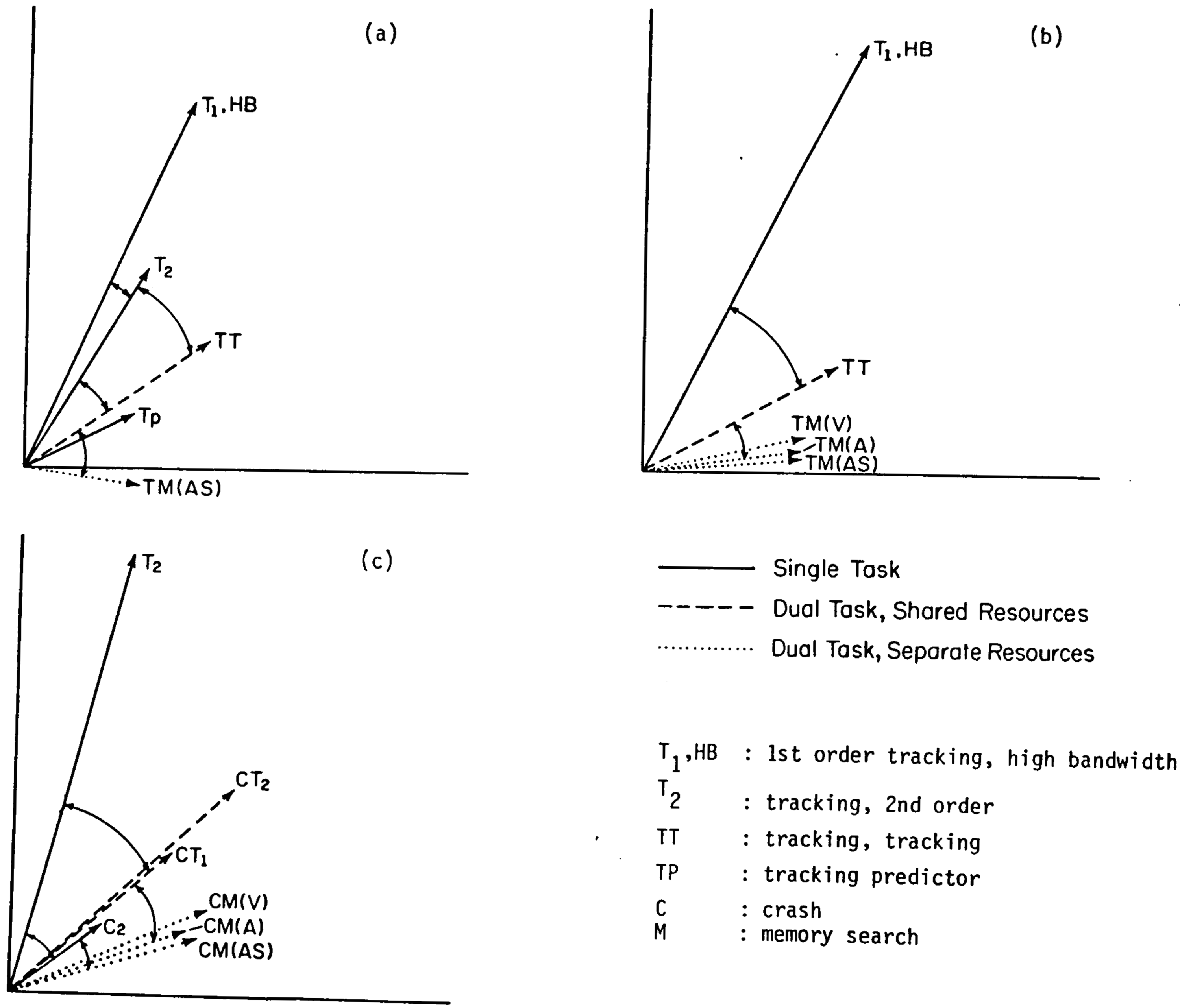


Figure 3: S-P Dissociations in Experiment 1 (a), 2 (b), and 3 (c)

the tracking task with a mixed order, nor the dual-axis tracking task with the integrated display). The effects of different manipulations on the standard task are characterized by vectors in the performance-subjective rating space with the origin representing the standard task condition. If two manipulations drive the two measures to the same degree (when normalized in terms of their variability), then the resulting vectors will lie along the same line. When two manipulations drive the two measures to different degrees, the resulting vectors will diverge.

ANOVAs were used to compare the effect of pairs of manipulations. This analysis is robust over the violation of the assumption of normality. Thus, a potential skewed distribution of either dependent measure will not cause extensive biases in the outcome. P and S were treated as two levels of one variable (i.e., measure type). The particular difficulty manipulation used was another variable. The interaction between the two variables is used to detect whether two difficulty manipulations drive the two measures to different degrees. If a pair of manipulations result in diverging vectors in the performance-subjective rating space, then the interaction will be be significant.

Using this analysis technique, the dissociation between S and P was found to be statistically reliable for the following pairs of vectors in Figure 3a.

(1) High bandwidth versus unaided second order tracking ($F(1,7) = 4.98$, $p < 0.1$). While the overall potency of the bandwidth manipulation was greater than that of the order manipulation (the vector is larger in both dimensions), on a relative basis bandwidth appears to drive P to a greater extent and order to drive S. This dissociation was not found when P was contrasted with the ratings on time-demand, input complexity, or response complexity scale. The lack of dissociation in these three scales

suggests that the differences in the P may be better indicated by ratings on either of the three scales rather than by overall workload ratings.

(2) Unaided second order tracking versus dual-axis tracking ($F(1,7) = 15.50$, $p < 0.01$). This is a "strong dissociaton" in the sense that the potency of the two manipulations are equivalent, but on an absolute basis, S (second order) < S (T&T) but P (second order) > P (T&T). The same kind of strong dissociation was also found in comparing bandwidth manipulation with dual-axis tracking manipulation, $F(1,7) = 32.70$, $p < 0.005$. This finding clearly replicates the dissociation earlier observed by Wickens and Derrick (1981): Increasing single task difficulty drives P more than S. Doubling the number of tasks drives S more than P.

(3) Dual-axis tracking versus tracking time-shared with memory search ($F(1,7) = 6.32$, $p < 0.05$). This effect is consistent with the results obtained by Wickens and Derrick (1981). That is, while both dual task conditions produce a substantial increment in perceived difficulty, the dual axis tracking task, because of the competition for processing resources also degrades performance. On the other hand, tracking and memory search, demanding separate resources are time-shared quite efficiently and actually generate a small net <u>gain</u> in total dual task performance, i.e., a time-sharing <u>increment</u>. Examination of the data in Table 7 reveals that this "increment" is produced by a small reduction in RT, and a small increase in tracking error.

(4) Second order tracking with versus without the predictor ($F(1,7) = 9.61$, $p < 0.05$). While performance was substantially assisted by the predictor element, subjective ratings were unaffected by its addition. This dissociation is manifest also when mixed order tracking is contrasted with second order tracking using the predictor. In Table 7, both tasks give equivalent performance, yet the prediction tracking trask is viewed

as subjectively more difficult.

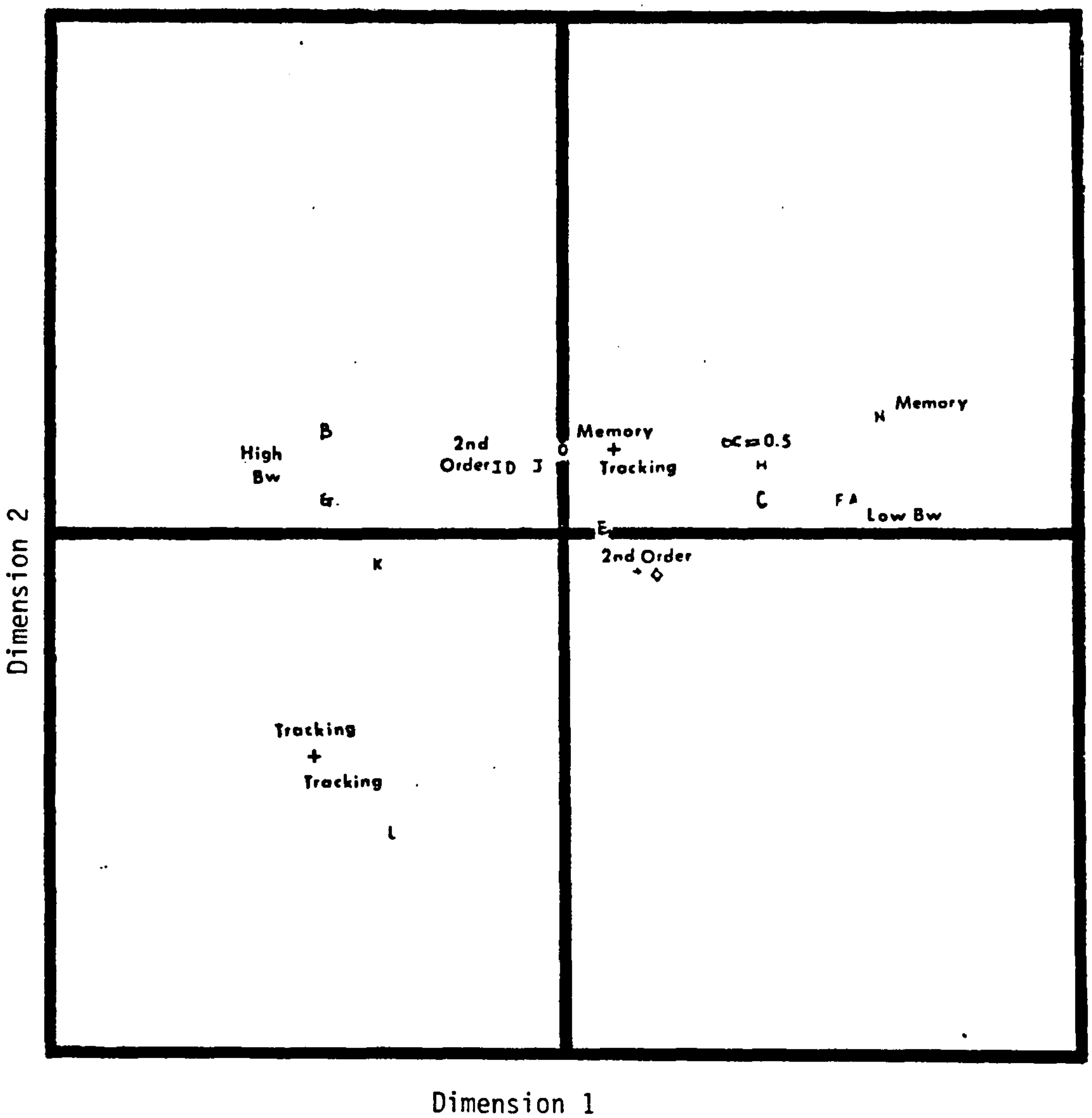
Similarity Judgment Analysis

Multidimensional Scaling (MDS) data were analyzed by the SINDSCAL (a version of INDSCAL) method. Three dimensions were chosen to represent the psychological structure underlying the difficulty ratings. The variance accounted for (VAF) by this three-dimensional solution was 61%. These dimensions are shown in Figures 4 and 5. Note that each single task configuration is represented twice, since different ratings were made for its performance with the left and right hand.

The correlations of these dimensions with the various unidimensional ratings are shown in Table 6. The first dimension was related to resource demand which is highly associated with all unidimensional ratings. Positive Dimension 1 weights were associated with tasks that demand less resources. Tasks that demand more resources (e.g., second order, high bandwidth, and dual-axis tracking) had negative weights (Figure 4).

Dimension 2 weights were most strongly correlated with unidimensional ratings on psychological stress, input complexity, and response complexity. This dimension was not orthogonal to the first dimension because the correlation between the two dimensions was 0.5. Examining the spatial representation, this dimension appears to be related to general resource competition, separating the two dual axis tracking configurations in which such competition was present, from the remainder of the tasks. In this regard, it is interesting to note that the location of the second order tracking task with the predictor is closer to the dual axis tracking tasks. This, like dual axis tracking, is the only other task in which processing resources must be shared between two elements (i.e., the cursor and predictor symbol).

The third dimension was not related to any unidimensional ratings.



A: 1st order, low bandwidth	(left hand),	F: 1st order, low bandwidth	(right hand
B: 1st order, high bandwidth	(left hand),	G: 1st order, high bandwidth	(right hand
C: mixed order	(left hand),	H: mixed order	(right hand
D: 2nd order, no predictor	(left hand),	I: 2nd order, no predictor	(right hand
E: 2nd order with predictor	(left hand),	J: 2nd order with predictor	(right hand

K: Dual-axis tracking in a separated display
L: Dual-axis tracking in an integrated display
N: Memory search
O: Tracking shared with a memory search task

Figure 4: SINDSCAL solution from multidimensional scaling data - Dimension 1 vs. Dimension 2 (Experiment 1)

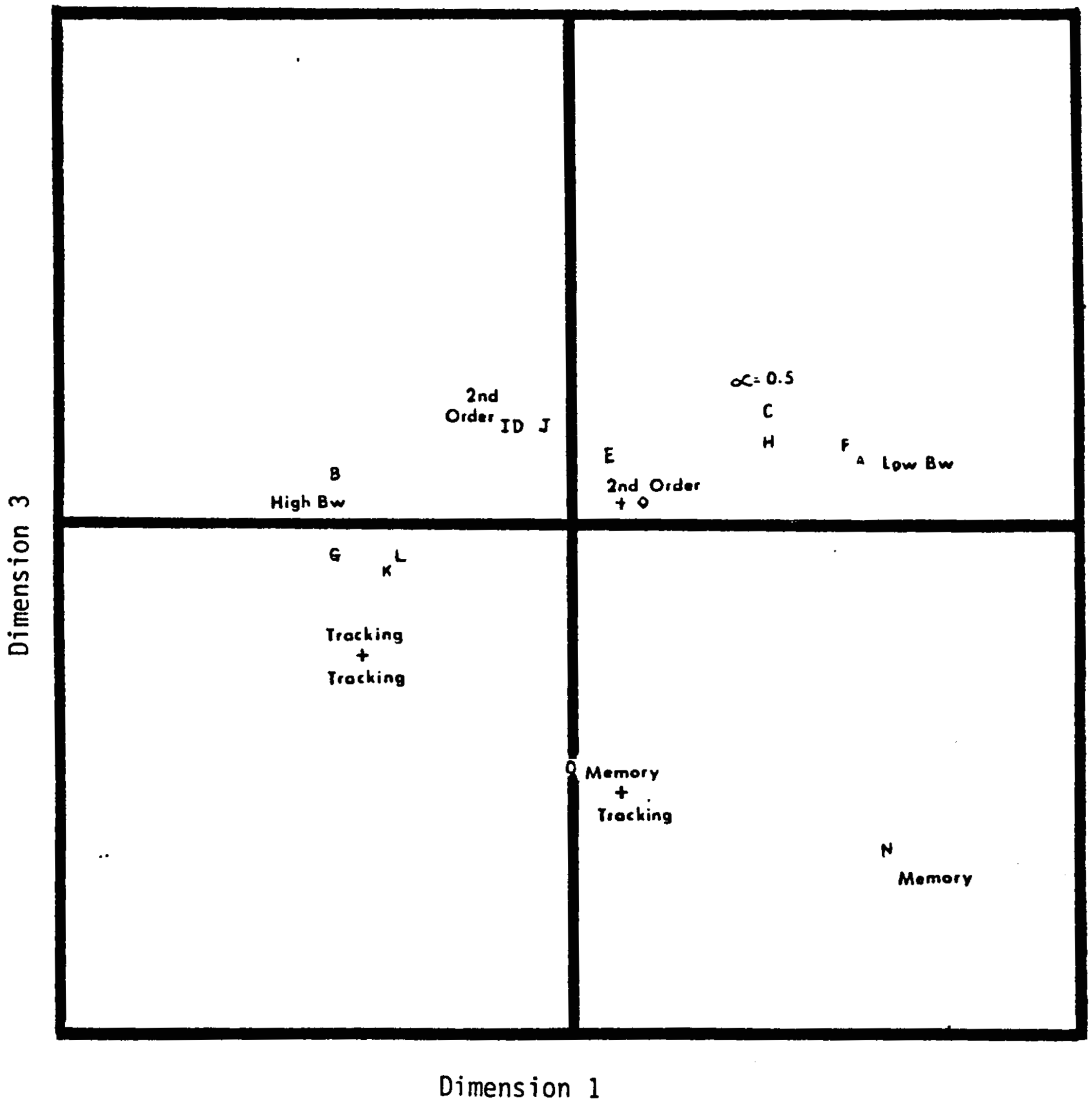


Figure 5: SINDSCAL solution from multidimensional scaling data - Dimension 1 vs. Dimension 3 (Experiment 1)

Examining the tasks in the psychological space, it is suggested that this dimension is associated with input/output modality and processing codes (Figure 5). Tasks that primarily involved the tracking task (visual-spatial-manual processing) located in the upper half of the dimension. Highly negative weights were associated with tasks that involved the memory search task (auditory-verbal-speech processing). Derrick (1982) had similarly uncovered a difficulty dimension associated with input modality.

The SINDSCAL technique also allows us to examine individual differences in weighting dimensions in judgment. According to the weights that each subject assigned to the three dimensions, there were pronounced individual differences in weighting the dimensions. Four subjects employed only one dimension to judge the similarity of task difficulty. The other four used various combinations of two dimensions.

Experiment 2

The effect of manipulating input/output modality of a Sternberg memory search task on the two measures, S and P, was investigated in this experiment. The purpose of introducing various input/output combinations is to induce different degrees of resource competition in dual task conditions when the task is time-shared with a tracking task (visual-manual). It has been found that the manipulation of input/output competition affects dual-task performance (Wickens, Sandry, & Vidulich, 1983). According to the hypothesis proposed above, varying amounts of I/O competition should not affect S to a great extent, so long as the number of tasks (2) remains constant.

Method

<u>Subjects</u>. Seven students who served as subjects in the first

experiment also performed in Experiment 2.

Sternberg memory search task. The four versions of the Sternberg tasks, i.e. visual-manual (VM), visual-speech (VS), auditory-manual (AM), and auditory- speech (AS) were used. With the speech response, subjects articulated their response into the microphone. With the manual response, they pressed one of two buttons for a positive or negative response. The task was responded with the left hand when the stimuli were presented on the left-visual-field or from the left ear. The task was responded to with the right hand when stimuli were input from the right side.

Tracking task. The low and high bandwidth first order tracking tasks were employed.

Dual task combinations. The four versions of the Sternberg memory search tasks were paired with the low-bandwidth tracking task. When the memory task was visually presented, it was shown on the left of the vertical tracking display or above the horizontal display. The dual-axis tracking with the separated display configuration was also employed in this experiment. The conditions employed in Experiment 2 are shown in Table 8).

Procedure. Subjects performed and rated all the tasks in the same way as they did during the first experiment. However, only interval ratings were collected in this experiment.

Results

Correlational analysis

(A) Test-retest reliability. As in Experiment 1, most subjects were more consistent on rating tasks at the end of a session. As shown in Table 9, the overall reliability of the ratings on the four attributes at the end of a trial was 0.81. The reliability of corresponding four ratings at the end of a session was 0.89. The overall reliability of

I. Single task conditions

(1) low bandwidth 1st-order tracking on the left hand.
(2) high bandwidth 1st-order tracking on the left hand.
(3) VM (visual-manual) memory search task (on the left side).
(4) VS (visual-speech) memory search task (on the left side).
(5) AM (auditory-manual) memory search task (on the left side).
(6) AS (auditory-speech) memory search task (on the left side).
(7) - (12) same as above presented on the right side of the display or to the the right ear.

II. Dual task conditions

(13) 1st-order Tracking (left side) with VM (right side)
(14) 1st-order Tracking with VS
(15) 1st-order Tracking with AM
(16) 1st-order Tracking with AS
(17) - (20) same as above except that tracking was on the right side and memory task was on the left side of the display.
(21) Dual-axis tracking

Table 8 - Experimental conditions (Experiment 2)

Scale	at the end of a trial	at the end of a session
Overall workload	0.81	0.92
Complexity	0.82	0.96
Psychological stress	0.76	0.85
Time-demand	0.84	0.83
Task demand		0.89
Input complexity		0.88
Mental effort		0.84
Response complexity		0.89
Rank order		0.96

(No of tasks = 21, all correlations are significantly greater than 0.0, $p < .01$)

Table 9 - Test-retest reliability (Experiment 2)

rating the eight attributes at the end of a session was 0.89. The averaged reliability of similarity judgments was 0.80.

(B) Correlation among unidimensional ratings and with performance. As in Experiment 1, the inter-scale correlations were quite high. Examining the original P and S scores shown in Table 10 reveals same fairly predictable trends. Subjective workload increased at high bandwidth and with all of the dual task conditions. Both reaction time and tracking error increased slightly in the dual-task conditions. However, the correlation between the accuracy/reaction time and the workload ratings is low in all memory search configurations. The correlation between tracking error and workload ratings is also low (Table 11). The tracking performance was correlated with the ratings on feedback adequacy, success of performance, and the nature of the task (Table 11).

Performance and Subjective Workload Analysis

Decrements from the baseline condition (low-bandwidth and the appropriate single task Sternberg configuration) were calculated and normalized in the same manner as in Experiment 1. Dissociations were found to be statistically reliable for the following pairs of vectors in Figure 3b.

(1) High bandwidth versus dual-axis tracking manipulation ($F(1,6) = 39.33$, $p < 0.005$). Comparing this figure with Figure 3a reveals a very close replication of the strong dissociation between these two manipulations. Subjects felt more loaded when performing dual-axis tracking task than a high bandwidth tracking task even though their performance was worse in the latter condition.

(2) Dual-axis tracking versus tracking time-shared with an AM memory task ($F(1,6) = 4.62$, $p < 0.10$) and tracking time-shared with an AS memory task ($F(1,6) = 5.44$, $p < 0.10$). Again, replicating the result of

	Accuracy	RT	RMS	Workload
Low-bandwidth Tr			0.14	2.64
High-bandwidth Tr			0.28	6.21
VM	96.59	0.74		3.07
VS	97.44	1.01		3.36
AM	95.88	1.28		3.00
AS	97.47	0.97		3.14
Tr + VM	95.92	0.80	0.15	6.36
Tr + VS	96.86	1.09	0.16	6.29
Tr + AM	95.99	1.29	0.15	5.93
Tr + AS	97.36	1.03	0.14	6.07
Tr + Tr			0.17	6.64

Table 10 - Original scores (Expriment 2)

	Wld	Comp	Strs	Time	TkDe	Inp	MnEf	Resp	Fb	Perf	Nat	Rank	Exce	Ac
Comp	.99													
Strs	.99	.99												
Time	.99	.97	.98											
TkDe	.98	.99	.99	.96										
Inp	.99	.99	.98	.97	.98									
MnEf	.93	.96	.92	.88	.94	.95								
Resp	.98	.97	.99	.97	.99	.98	.90							
Fb	.81	.86	.86	.78	.86	.82	.79	.83		(all are significant)				
Perf	.94	.95	.97	.90	.97	.94	.89	.96	.90					
Nat	.96	.96	.96	.94	.95	.94	.88	.95	.87	.96				
Rank	.95	.87	.93	.79	.95	.77	.77	.94	.78	.97	.87			
Exce	.98	1.0	.98	.96	.99	.99	.97	.97	.86	.95	.95	.88		
MDS dimensions									("-"'s are not significant)					
D1	.98	.95	.95	.97	.94	.97	.79	.94	.76	.88	.92	.95	.97	
D2	.57	.57	.65	.48	.50	.59	.10-	.58	.37-	.47	.57	.59	.44	
Tracking Error									("*"'s are significant)					
Rms	.35	.32	.46	.35	.40	.26	.08	.47	.66*	.58*	.57*	.47	.30	
Memory Performance								(none is significant)						Ac
Acc	.12	.16	.14	.17	.13	.19	.14	.08	.38	.08	.09	.05	.15	
RT	.11	.10	.09	.09	.14	.18	.09	.12	.15	.15	.13	.12	.11	-.15

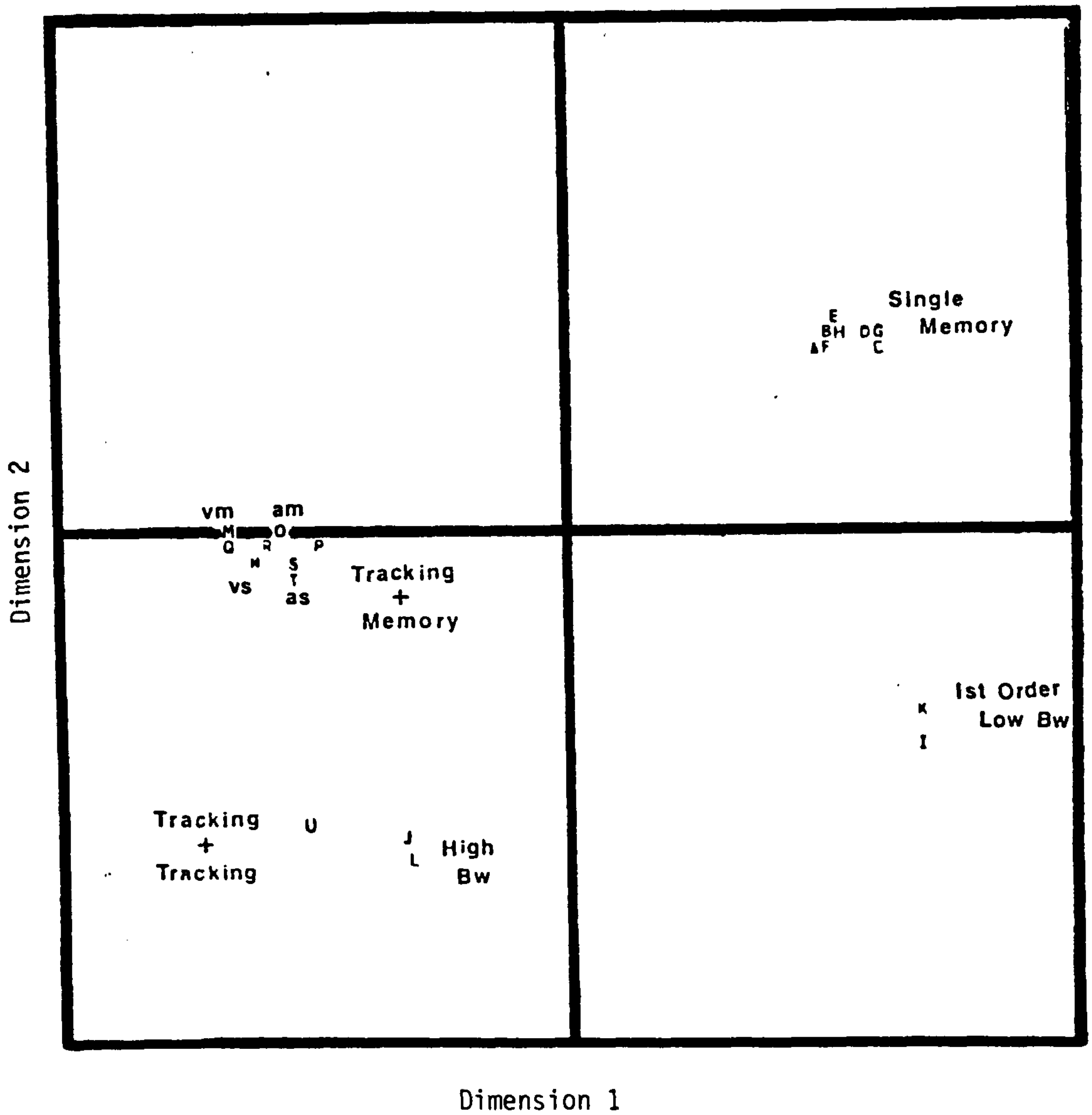
Table 11 - Correlation among unidimensional ratings , Performance, and MDS dimensions (Experiment 2)

Experiment 1, performance rather than subjective ratings were driven more by the competition for common resources in the input modality pool.

The competition for output resources did not show any effect on either P or S. The competition for input resources affected performance. Subjects performed better when the two tasks distributed demand over visual and auditory resources than when both tasks demanded visual resources ($F(1,6)=9.59$, $p < 0.05$). But, subjective ratings did not show such an effect. However, the interaction in the dissociation ANOVA was not statistically significant. The proximity of all vectors to each other suggests in general that there were little substantial effects of input/output modality competition on either dependent variable.

Similarity Judgment Analysis

Two dimensions were disclosed by the SINDSCAL method and are plotted in Figure 6. As in Figure 4, each single task configuration is represented twice since different ratings were made for its performance with the left and right hand. The VAF by the two-dimensional solution was 0.69. Dimension 1 was highly correlated with every unidimensional ratings (Table 11). This (horizontal) dimension is related to resource demand. Highly positive weights were associated with tasks that demand less resources. Negative weights were related to tasks that demand more resources. Dimension 2 was correlated with ratings on psychological stress, task difficulty, input complexity, and response complexity. This dimension was not orthogonal to the first dimension since the correlation between the two dimensions was 0.48. Locations of the tasks in the spatial representation indicate that this dimension is related to processing codes. Positive weights, zero weights, and slightly negative weights were associated with memory search tasks (verbal codes). Negative weights were associated with configurations that involved tracking tasks



A: VM (left) B: VS (left) C: AM (left) D: AS (left)
E: VM (right) F: VS (right) G: AM (right) H: AS (right)
I: 1st order, low bandwidth (left) J: 1st order, high bandwidth (left)
K: 1st order, low bandwidth (right) L: 1st order, high bandwidth (right)

M: VM + tracking N: VS + tracking
O: AM + tracking P: AS + tracking

Q: tracking + VM R: tracking + VS
S: tracking + AM T: tracking + AS

U: tracking + tracking

Figure 6: SINDSCAL solution from multidimensional scaling data - Dimension 1 vs. Dimension 2 (Experiment 2)

(spatial codes). As in Experiment 1, there were individual differences in weighting the two dimensions. Three subjects primarily used the first dimension to judge the similarity. Two subjects used the second dimension and the other two subjects used both dimensions to judge the similarity of task difficulty.

Experiment 3

In this experiment, the "Crash" task was employed with the tracking and the memory search task. The crash task and the first order tracking task were used as the baseline condition from which the difficulty was manipulated. As in the previous two experiments, demands upon various components of processing resources were systematically imposed by different task combinations. The purpose of employing the crash task was to replicate and generalize results from the previous two experiments to a different spatial task.

Method

Subjects. Eight subjects participated in this experiment. Five of them had served as paid volunteers in the first two experiments.

Crash task. Subjects responded to this task manually or with speech. They were instructed to enter their judgments on a five-key keyboard or to speak their answers as quickly and as accurately as possible. When the answers were spoken, subjects were told to leave a slight pause between the two judgments in order to avoid the limitation of speech recognition unit.

The difficulty of the task was manipulated by the degree of perceptual discrimination in the second judgment. This task was always performed with the right hand.

Tracking task. The dynamics of a vertical tracking task was either first or second-order. This task was always performed with the left hand.

Sternberg memory search task. The four versions of the Sternberg task (i.e., VM, VS, AM, and AS) were employed. Subjects either pressed a key with the left hand or vocalized their responses.

Dual task. The easy crash task was paired with either the first-order, or the second-order tracking task. Subjects responded to the crash task either manually or with speech in these two conditions. The easy crash task was also performed with one of the four versions of the memory search task. In these four dual-task conditions, subjects always responded manually to the crash task. A summary of all of the experimental conditions is shown in Table 12.

Procedure. Subjects were trained on the crash task during the first two days. In these sessions, five separation ranges were randomly chosen within a two-minute trial. Feedback on the two decisions was displayed at the end of intercept incidence during a trial. After these two training sessions, subjects practiced all the tasks for another three days. Accuracy of the first decision and the proximity of the second decision (correlation between the actual and subject's estimation) were presented to the subject at the end of each trial. During the last three sessions, subjects performed and rated all tasks in the same way as they did in the second experiment. At the end of the last session, subjects were asked to rate the importance of the 19 factors shown in Table 4 to their definition of mental workload.

Results

Correlational Analysis

(A) Test-retest reliability. As in the previous two experiments, most subjects were more consistent on rating tasks at the end of a session than at the end of a trial. The overall reliability of rating four unidimensional attributes was 0.70. The overall reliability of

I. Single task condition

(1) Easy Crash task with manual response - Cr (m)
(2) Easy Crash task with speech response - Cr (s)
(3) Difficult Crash task with manual response
(4) Difficult Crash task with speech response
(5) 1st-order tracking
(6) 2nd-order tracking
(7) VM
(8) VS
(9) AM
(10) AS

II. Dual task conditions

(11) 1st-order tracking with Cr (m)
(12) 1st-order tracking with Cr (s)
(13) 2nd-order tracking with Cr (m)
(14) 2nd-order tracking with Cr (s)
(15) VM + Cr (m)
(16) VS + Cr (m)
(17) AM + Cr (m)
(18) AS + Cr (m)

Cr(m) - Crash task with manual response
Cr(s) - Crash task with speech response
Tr 1 - 1st-order tracking task
Tr 2 - 2nd-order tracking task
VM - visual-manual version of the memory task
VS - visual-speech
AM - auditory-manual
AS - auditory-speech

Table 12 - Experimental conditions (Experiment 3)

corresponding ratings at the end of a session was 0.89. The overall reliability of rating the eight attributes at the end of a session was 0.89. The averaged reliability of similarity judgment was 0.76 (Table 13).

(B) Correlation among unidimensional ratings and with performance. As in previous experiments, the inter-scale correlation of the mean ratings across subjects were quite high (Table 14). Once again, these high correlations suggest that subjective workload may reflect a summation demand of every aspect of information processing.

Examining the raw performance and subjective ratings in Table 15, it is clear that objective and subjective workload was increased by all difficulty manipulations. Tracking error increased in the second-order tracking task and when tracking was performed with the easy crash task. Also, the difference between first and second order tracking was enhanced in the dual task conditions. However, as shown in Table 14, RMS error was not significantly correlated with any undimensional ratings.

The accuracy of the first judgment of the crash task did not decrease in the difficult or the dual-task condition, nor were the mean accuracy data correlated with any unidimensional ratings. The proximity judgment of the second decision decreased in the difficult crash task, but not in the dual-task conditions. This performance was not related to workload rating, but was associated with the ratings of feedback adequacy. Reaction time of the crash task increased in the difficult as well as in the dual-task conditions. This performance was marginally correlated with ratings on task demand, task difficulty, and psychological stress.

Reaction time, but not accuracy of the memory search task changed significantly when the memory task was paired with the easy crash task. Neither performance measure was significantly correlated with any unidimensional ratings.

Scale	at the end of a trial	at the end of a session
Overall workload	0.76	0.87
Complexity	0.85	0.92
Psychological stress	0.55	0.90
Time-demand	0.72	0.86
Task demand		0.84
Input complexity		0.91
Mental effort		0.88
Response complexity		0.88
Rank order		0.92

(all are significant at $p < 0.01$)

Table 13 - Test-retest reliability (Experiment 3)

	Wld	Comp	Strs	Time	TkDe	Inp	MnEf	Resp	Fb	Perf	Nat	Rank	Exce	Ac	Cor
Comp	.99														
Strs	.99	.99													
Time	.99	.99	.99												
TkDe	.99	.99	.99	.99											
Inp	.99	.99	.98	.99	.99										
MnEf	.96	.97	.97	.96	.97	.98									
Resp	.99	.97	.98	.99	.99	.98	.94								
Fb	.63	.71	.62	.66	.68	.68	.73	.61							
Perf	.95	.96	.93	.95	.94	.96	.93	.92	.77	(all are significant)					
Nat	.97	.99	.96	.98	.97	.98	.95	.97	.74	.99					
Rank	.95	.97	.96	.97	.98	.97	.93	.97	.73	.96	.97				
Exce	.99	.98	.99	.99	.99	.98	.96	.99	.63	.94	.97	.96			
MDS dimensions										(all are significant)					
D1	.95	.95	.97	.94	.95	.96	.97	.93	.60	.88	.91	.96	.95		
D2	.59	.60	.53	.63	.58	.60	.48	.63	.53	.70	.70	.53	.61		
Crash Performance	("*",s are significant)													Ac	Cor
Ac	.23	.21	.23	.22	.21	.25	.23	.27	.04	.25	.26	.20	.23		
Cor	.01	.12	.06	.01	.01	.03	.01	.07	.61	.22	.19	.09	.09	-.01	
RT	.47	.49	.51*	.48	.52*	.44	.46	.47	.18	.33	.42	.55*	.47	-.55*	-.11
Tracking Error										(none is significant)					
Rms	.28	.12	.10	.20	.11	.23	.08	.35	.62	.11	.21	.13	.20		
Memory Performance										(none is significant)				Ac	
Ac	.27	.45	.27	.48	.38	.56	.25	.38	.15	.31	.33	.38	.28		
RT	.13	.34	.18	.41	.30	.37	.11	.46	.48	.34	.42	.35	.21	-.56	

Table 14 - Correlation among unidimensional ratings , Performance, and MDS dimensions (Experiment 3)

	RMS	Crash Performance			Memory Performance		Workload
		Acc	Corr	RT	Acc	RT	
1st-order Tr	0.16						3.29
2nd-order Tr	0.29						5.43
Easy Cr(m)		99.33	0.89	3.27			3.29
Easy Cr(s)		95.56	0.89	3.38			3.29
Difficult Cr(m)		99.02	0.75	3.48			4.29
Difficult Cr(s)		95.10	0.74	3.53			4.71
Tr 1 + Cr(m)	0.27	99.15	0.88	3.33			5.71
Tr 1 + Cr(s)	0.25	95.37	0.86	3.58			5.71
Tr 2 + Cr(m)	0.43	97.71	0.83	3.34			7.86
Tr 2 + Cr(s)	0.42	92.47	0.88	3.65			8.00
VM					97.39	0.90	2.71
VS					96.84	1.13	2.71
AM					97.28	1.36	2.57
AS					95.93	1.08	2.57
VM + Cr		97.86	0.89	3.56	95.77	1.22	6.29
VS + Cr		98.98	0.85	3.57	94.08	1.48	6.43
AM + Cr		96.30	0.87	3.59	95.48	1.65	5.71
AS + Cr		97.07	0.87	3.60	96.85	1.33	5.71

Table 15 - Original scores (Experiment 3)

Performance and Subjective Workload Analysis

Original scores are presented in Table 15. Decrement scores were computed from the appropriate single task configuration and normalized for each dependent variable following the procedures outlined above. These data are presented in Figure 3c. Dissociations were found in the following comparisons:

(1) Second-order tracking versus tracking time-shared with an easy crash task ($F(1,7) = 17.11$, $p < 0.01$). As was found in both Experiments 1 and 2, a strong dissociation was shown in contrasting difficulty manipulation via the single task or through adding another task. Subjects' performance decrements were lower but subjective workload was higher when time sharing a first-order tracking task with an easy crash task, than when performing a second-order tracking task. This dissociation, however, was not replicated in the comparison between the difficult crash and the crash task time-shared with the tracking task.

(2) Second-order tracking versus difficult crash task ($F(1,7) = 12.61$, $p < 0.01$). With equal level of subjective workload, performance in the difficult tracking task condition was worse than that in the difficult crash task condition. Stated differently, increasing the complexity of the crash task had a relatively greater impact on subjective ratings (per given loss in performance), than increasing the order of tracking.

(3) Difficult crash versus crash time-shared with memory search task ($F(1,7) = 11.39$, $p < 0.05$). Performance decrements were approximately equivalent under these two manipulations while subjective ratings were significantly higher in the dual-task condition.

(4) Crash task (manual) time-shared with first-order tracking versus crash (manual) time-shared with a VM memory search task ($F(1,7) = 8.84$, $p < 0.05$), a VS memory search task ($F(1,7) = 5.24$, $p < 0.1$), a AM memory search task ($F(1,7) = 16.4$, $p < 0.015$), and a AS memory search task ($F(1,7) = 8.184$, $p < 0.025$). Performance

decrements were higher when the crash and tracking task were time-shared competing for resources of spatial codes, than when the crash and memory task were shared, distributing the demand over spatial and verbal codes. However, subjective workload was not affected by the amount of resource competition.

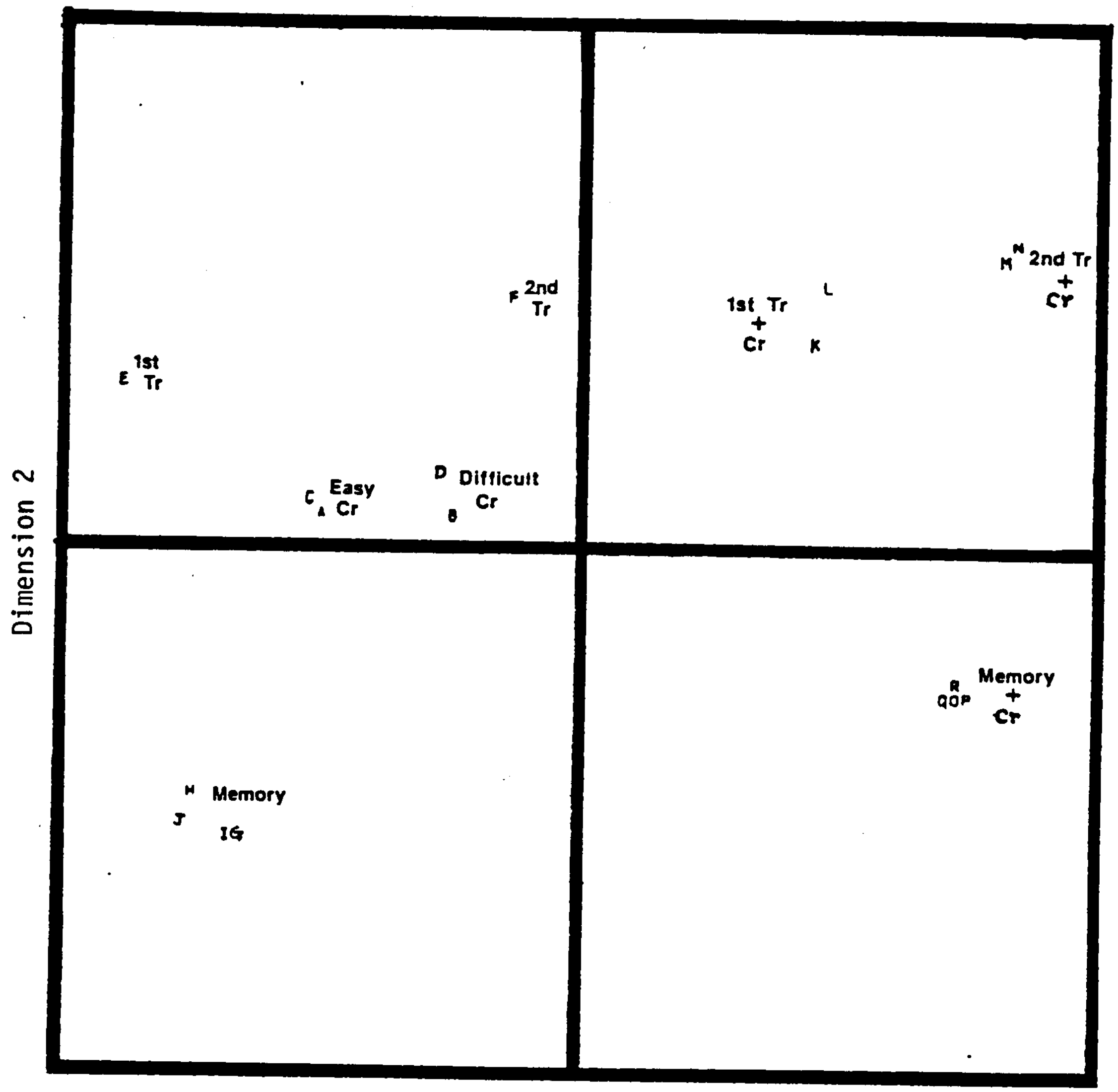
Two dimensions were selected to represent the psychological structure of all tasks. The VAF by this two-dimensional solution was 65%. The first dimension was related to the resource cost. Negative weights were associated with tasks that demand fewer resources. Positive weights were related to tasks that demand more resources (Figure 7). As in Experiments 1 and 2, the second dimension was associated with processing codes. Negative Dimension 2 weights were associated with configurations that included a memory task (verbal codes). Configurations that included spatial codes (tracking or crash task), were located in the upper half of the space. The correlation between the two dimensions was 0.36. Thus, the two dimensions were not completely orthogonal. Individuals were also different in weighting these two dimensions. Three subjects primarily used the first dimension, two used the second dimension, and the others used both dimensions to judge the similarity of task difficulty.

Importance of 19 Factors to Subjects' Definition of Mental Workload

Among the 19 factors, the following ones were chosen as a primary element by most subjects: (1) task demand-amount, (2) fatigue, (3) environment, (4) performance, and (5) emotional stress. These five factors had high factor loadings in the seven factors suggested by Hart et al. (1982).

Discussion

In summary, data from the three experiments showed the following



A: Easy crash (manual response)
C: Easy crash (speech response)

B: Difficult crash (manual response)
D: Difficult crash (speech response)

E: 1st order tracking

F: 2nd order tracking

G: VM
I: AM

H: VS
J: AS

K: 1st order + easy crash (manual)
M: 2nd order + easy crash (manual)

L: 1st order + easy crash (speech)
N: 2nd order + easy crash (speech)

O: VM + easy crash (manual)
Q: AM + easy crash (manual)

P: VS + easy crash (manual)
R: AS + easy crash (manual)

Figure 7: SINDSCAL solution from multidimensional scaling data - Dimension 1 vs. Dimension 2 (Experiment 3)

results:

(1) Subjects are consistent and reliable on their workload ratings. Although both post trial and post session reliabilities were high, most subjects were more reliable when rating a task at the end of a session than at the end of a trial. When subjects rate a task at the end of a trial, they may introspect the immediate experience. The immediate experience may fluctuate whenever there is a change in the effort investment, motivation, or the difficulty of a particular memory set. When subjects rate a task at the end of a session, the task is rated in the context of all tasks performed and is therefore more of a relative than an absolute judgment.

(2) There is a high correlation between the transformed ratio and interval ratings. This result suggests that these two rating methods provide essentially the same results. The correlations among ratings on unidimensional attributes are also very high. These high correlations suggest that subjective workload may reflect the combined estimates of demand on every aspect of information processing.

(3) Replicating Derrick's (1984) findings, a robust dissociation is found by manipulating task difficulty via the single task (high bandwidth, second order tracking, or difficult crash task) or through adding another task (dual-axis tracking, tracking time-shared with memory or crash task or predictor tracking including the sharing of spatial resources between two display elements). This dissociation indicates that P is driven more by increasing the difficulty of a single task while S is driven relatively more by the number of concurrent tasks. This dissociation is observed across different manipulations in the three experiments.

(4) A second robust dissociation is observed by manipulating resource competition. Results suggest that P is driven more by competition between

tasks for common resources (e.g., dual-axis tracking, tracking with crash), whereas S appears to be relatively insensitive to the degree of resource competition between tasks. Simply time-sharing two tasks, whether common or separate resources are demanded, is sufficient to create a high level of subjective workload. The only instance where this dissociation was only weakly observed occurred in Experiment 2. Separate comparisons revealed that performance distinguished perceptual competition, while subjective ratings did not. However, this dissociation was not revealed by the statistical interaction, nor was any evidence of a dissociation for output competition obtained.

(5) A weak dissociation is observed by increasing difficulty that imposes response load versus increasing difficulty by imposes perceptual/cognitive load. In the first experiment, this effect is observed by manipulating control order versus bandwidth. From previous dual-task studies, it has been shown that control order imposes demand on perceptual/cognitive as well as response load while bandwidth increases response load only (Wickens, 1976; Wickens & Derrick, 1980; Isreal, Chesney, Wickens, & Donchin, 1980). Thus, subjects felt more loaded when performing a second-order tracking task which increases perceptual/cognitive demand when performing a high bandwidth tracking task which imposes only on response resources. Note that such a dissociation was not found when performance was contrasted against ratings on input complexity, response complexity, or time-demand. A second apparent manifestation of this dissociation was found in Experiment 3 in which the second order task, demanding perceptual <u>and</u> response related resource, showed a relatively smaller influence on S, than did the difficult crash task, whose added demands were almost exclusively perceptual.

The above three dissociation effects confirm two predictions of the theory: (i) Parameters of a single task influence P more than S and this difference is more pronounced if the task is degraded by imposing demands on responding, rather than on perceptual/cognitive processing. (ii) The number of concurrent tasks increase S and decrease P, but the former by a greater degree than the latter.

(6) When subjects perform a second-order tracking task with a predictor, their performance is facilitated while their overall workload is not reduced. One interpretation of this dissociation is that S is driven by the number of separate display elements just as it is by the number of task. This increase in number when the predictor is present compensates for any reduction in perceived difficulty that might result from the useful lead information provided by the predictor. A second interpretation is that the predictor, by providing more useful stimulus information actually leads the subject to invest <u>more</u> resources into the task. While greater resource investment may help, and certainly does not harm performance, it will lead to greater values of <u>S</u>. Results consistent with this interpretation were obtained by Vidulich (1983), who found that increased incentives to perform well on a tracking task increased performance, but also increased subjective workload.

(7) The underlying structure of subjective perception of task difficulty is related to resource cost. Two dimensions are consistently found from the MDS analysis across three experiments. One dimension is associated with the total resource cost, and a second is related to the processing codes. In addition, in Experiment 1 a third dimension seemingly related to resource competition was found. These results converge with Derrick's findings, in revealing that the underlying structure of subjective workload is not one-dimensional, and may be partially interpreted within the framework of the multiple resource

theory. Furthermore, there are individual differences in weighting each dimension in perceiving task difficulty.

Systematic research is needed to understand exactly how subjective workload relate to every resource dimension in the multiple resource model and to verify the theory proposed in this paper. Nevertheless, the dissociation effects found in this study indicate a limitation of predicting performance from subjective ratings of workload alone. Even though subjective workload may indeed represent a valuable tool to facilitate system design, this measure may not indicate some aspects of human information processing which are important to system performance and safety. The strong dissociation resulted from manipulating difficulty of a single task suggests that designers can choose a non-optimal system because a bias. When designers are guided by the criterion of minimizing subjective ratings of mental workload in their selection of optimal systems, they will be biased in their choice away from those systems that may have multiple tasks, and may be blinded to the particular advantages to performance resulting for separate resources. In contrast, they will be biased toward those configurations that involve single tasks, even though the latter may produce relatively poorer performance. This bias may be more pronounced if the single tasks are heavily response loaded. Since system <u>performance</u> is the ultimate criterion against which systems must be judged, it is vitally important that the user of subjective ratings be aware of the biases against certain generic classes of tasks that may be induced.

References

Ashkenas, I. L. Some open and closed loop aspects of airplane lateral-direction handling qualities. Paris: NATO, AGARD Dograph No. 533, 1966.

Borg, G. Subjective effort in relation to physical performance and working capacity. In H. Pick (Ed.), Psychology from research to practice. New York: Plenum Press, 1978.

Casali, J. G., & Wierwille, W. W. A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activity. IEEE Transactions on Systems, Man, and Cybernetics, 1982.

Derrick, W. The relationship between processing resource and subjective dimensions of operator workload. Proceedings, 25th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors, 1981.

Derrick, W. A multiple processing resource explanation of the subjective dimensions of operator workload. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1984.

Ericksson, K., & Simon, H. Verbal reports as data. Psychological Review, 1980, 87, 215-250.

Eggemeier, F. T., Crabtree, M. S., & Reid, G. B. Subjective workload assessment in a memory update task. Proceedings, 26th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors, 1982.

Gopher, D., & Braune, R. On the psychophysics of workload: Why bother with subjective measures? Proceedings, Second Symposium on Aviation Psychology. Ohio State University, Columbus, OH, April, 1983.

Hart, S. G., Childress, M. E., & Hauser, J. R. Individual definitions of the term "workload". Paper presented at the 1982 Psychology in the DOD Symposium.

Hauser, J. R., Childress, M. E, & Hart, S. G. Rating consistency and component salience in subjective workload estimation. Proceedings, 18th Annual Conference on Manual Control, U.S. Government Printing Office, 1982.

Herron, S. A case for early objective evaluation of candiate display formats. Proceedings, 24th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors, 1980.

Hess, R. A. Prediction of pilot opinion ratings using an optimal pilot model. Human Factors, 1977, 19, 459-475.

Hicks, T. G., & Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors, 1979, 21, 129-143.

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology, 1980, 17, 259-273.

Johanssen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., & Wickens, C. D. Final report of experimental psychology group. In N. Moray (Ed.), Mental workload: Its theory and measurement. New York: Plenum Press, 1979.

Moray, N. Subjective mental workload. Human Factors, 1982, 24, 25-40.

Rahimi M., & Wierwille, W. W. Evaluation of the sensitivity and intrusion of workload estimation technique in piloting tasks emphasizing mediational activity. IEEE Transactions on Systems, Man, and Cybernetics, Seattle, WA, 1982.

Tulga, M. K. Dynamic decision making in multi-task supervisory control. M.I.T. Man-Machine Systems Laboratory Technical Report, September, 1978.

Tulga, M. K., & Sheridan, T. B. Dynamics decisions and workload in multitask supervisory control. IEEE Transactions on Systems, Man, & Cybernetics, 1980.

Vidulich, M. D., & Wickens, C. D. Processing phenomena and the dissociation between subjective and objective workload measures. University of Illinois Engineering Psychology Laboratory Tech. Report EPL-83-2/ONR-83-2, October, 1983.

Wickens, C. D. The effects of divided attention on information processing in tracking. Journal of Experimental Psychology: Human Perception & Performance, 1976, 2, 1-13.

Wickens, C. D.. The structure of attentional resources. In R. Nickerson, and R. Pew (Eds.), Attention and performance VIII. Englewood Cliffs, N J: Erlbaum, 1980.

Wickens, C. D. Processing resources in attention. In R. Parasuraman, and R. Davies (Eds.) Variety of attention. New York: Academic Press, 1984.

Wickens, C. D., & Derrick, W. Workload measurement and multiple resources. IEEE Transactions on Systems, Man, & Cybernetics, October, 1981.

Wickens, C. D., Mountford, S. J., & Schreiner, W. Multiple resources, task hemispheric integrity, and individual differences in time-sharing. Human Factors, 1981, 23, 211-230.

Wickens, C. D., & Yeh, Y. Y. The dissociation of subjective ratings and performance. IEEE Transactions on Systems, Man, & Cybernetics, Seattle, WA, 1982.

Wickens, C. D., & Yeh, Y. Y. The dissociation between subjective workload and performance: A multiple resource approach. Proceedings, 27th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors, 1983.

Wierwille, W. W., & Connor, S. A. Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. <u>Human Factors</u>, 1983, <u>25</u>, 1-16.

Wierwille, W. W., & Gutmann, J. C. Comparison of primary and secondary task as a function of simulated vehicle dynamics and driving conditions. <u>Human Factors</u>, 1978, <u>20</u>, 233-244.